SCIENCE CHINA
Life Sciences

SPECIAL TOPIC
• EDITORIAL •

February 2013 Vol.56 No.2: 101–103
doi: 10.1007/s11427-013-4436-x

# Next-generation sequencing technologies for personalized medicine: promising but challenging

CHEN Geng & SHI TieLiu*

*Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China*

Received January 5, 2013

**Citation:** Chen G, Shi T L. Next-generation sequencing technologies for personalized medicine: promising but challenging. Sci China Life Sci, 2013, 56: 101−103, doi: 10.1007/s11427-013-4436-x

In the past several years, next-generation sequencing (NGS) technologies have greatly revolutionized our approaches to explore and depict the characteristics and functions of the genomes for various species. The NGS technologies have been broadly used in diverse fields including genomics (genome sequencing and exome sequencing) [1,2], transcriptomics (RNA-Seq) [3,4] and epigenomics (ChIP-Seq, MeDIP-Seq, MBD-Seq, MRE-Seq and etc) [5–7], and a host of important findings have been achieved in these areas. *Science China Life Sciences* also has published a number of intriguing papers related to the NGS technologies in recent years [8–15]. Generally, if the studied organism does not have available reference genome, one can sequence its DNAs and then construct the genome using *de novo* assembly. Furthermore, the genome and exome sequencing enable researchers to identify and characterize the SNPs (single nucleotide polymorphisms), indels (insertions and deletions), inversions and other structural variations of interested species at single nucleotide resolution to investigate the underlying mechanisms of certain phenotypes or diseases. Using the whole transcriptome sequencing (RNA-Seq) technology, researchers have the opportunities to comprehensively inspect the transcriptional events and expression profiles of genes in cells [11,16]. Many applications have been carried out with RNA-Seq, such as exon-exon splice junction detection, alternative splicing identification, gene and isoform expression quantification, gene fusion inference and etc. In particular, RNA-Seq allows for reconstructing the transcriptome and identifying novel genes/isoforms of particular organism as well, which cannot conduct with microarrays. To study the epigenetic marks responded to environmental and developmental signals, NGS technologies can also be employed to profile the histone modifications (ChIP-Seq) and DNA methylations (MeDIP-Seq, MBD-Seq and MRE-Seq) in normal and disease states. Since NGS has the unsubstitutable advantages, many laboratories have adopted this technology for various researches. However, many biologists have not realized that the sequencing depth is directly associated with the number of detectable genes and SNPs, as well as other gene features, different experiments with different goals need to use distinct NGS strategies to generate related data. Xiao et al. [17] have summarized the applications with different depths, which provide a useful reference for biologists to design their related experiments. Meanwhile, the External RNA Control Consortium (ERCC) data has been proposed as internal control to monitor the consistence and reproducibility of different RNA-Seq data [18], Shi LeMing's group [19] has evaluated the usages of ERCC between different experiments for single organism and between different organisms, they found that the gene expression measurement highly depends on the RNA enrichment protocols (Poly(A) selection and RiboZero), which imply that the integration of multiple gene expression data from different RNA-Seq protocols should take cautions. The internal control is very important for monitoring the consistence between different NGS experiments, especially for the investigation of single cell expression profile. Otherwise, the technical artifacts and truly biological variations between different single cells will not be distinguished.

Personalized medicine refers to the right drug for the

*Corresponding author (email: tieliushi01@gmail.com)

right person at the right time with right dose. The first step for realizing personalized medicine is to decipher the disease related and drug sensitive genes and SNPs at multi-levels, and then carry out massive genetic screening for identifying those variations on each individual genome. For this purpose, NGS has enormously accelerated the advance of personalized medicine [20,21] and has been utilized to identify those variations of human genome and epigenome, including those variations for diseases. Together with other experimental analyses, it enables researchers to further uncover the causative mechanisms of genotype to phenotype. Through the exploration on the relevant NGS data and other related data, biomarkers for diagnosis, prognosis and therapy are being explored for different diseases. The related new findings will substantially improve patient treatments and facilitate drug developments. However, different individuals may have distinct response to the same drug owing to the variations among their genomes and/or epigenome. To ensure the drug safety and efficacy, it is necessary to screen the genetic variations for each individual. The advance in NGS technology makes it become faster, cheaper and more accurate approach for us to obtain a panoramic view of the variations among individuals; it also promises to accurately diagnosis and track the patients' response to drugs in therapy, and help to optimize the drug usage to reduce the medication side effect. In this special issue, several papers have touched this topic and discussed the application of NGS on disease studies. Chen's group discuss the ocular gene therapies and the identification of the related causative genes for molecular diagnosis with NGS for retinal disorders [22]; Ning and his colleagues [23] summarize the progress of Abacavir-induced hypersensitivity reaction study and discuss the potential application of NGS in translational pharmacogenetics field for personalized medicine.

Although the NGS technologies benefit a variety of studies, many challenges remain to be resolved to accurately and effectively process and interpret the NGS data. The challenges of analyzing NGS data come from the limitations of both NGS technologies and the correlated bioinformatics software. One of the challenges is the sequencing accuracy. Although NGS technologies can generate significant amount of data compared with Sanger's method, their sequencing accuracy is lower by 2–3 orders of magnitude than the latter one. Also each NGS platform currently on market has its own inherent biases with special error patterns which have great impacts on the data interpretations, such as SNP identification, SNV and indel detection. For this reason, we believe that many reported SNPs identified by the NGS technologies are false results and need to be further verified. To evaluate the effects and the quality of NGS technologies and the analytic bioinformatics pipelines for realizing personalized medicine, FDA (U.S. Food and Drug Administration, http://www.fda.gov/) initiated the SEQC (Sequencing Quality Control) project to assess the technical performance of different NGS technologies and the advantages and limitations of diverse bioinformatics strategies. Meanwhile, to facilitate the innovation and reduce the cost of the NGS technologies, NIH also launched the challenge grants for the technology development several years ago, with the hope that the cost of whole genome sequencing will be under 1000 USD in the near future [24]. Ultimately the low cost will allow sequencing the genome of an individual as a routine medical test and making the personalized medicine come true in reality.

Compared with Sanger sequencing method, NGS vastly increases the output with significant cost and time effectiveness for sequencing, however, the reads of NGS are largely shorter than the ones from original Sanger methods in length with relatively higher sequencing errors in the reads. Furthermore, the huge amount of NGS data greatly raise the sequencing coverage, but more computational resources and more effective bioinformatics algorithms are needed to smoothly process the colossal dataset. In addition, the short length of NGS reads poses the challenge for short-read mapping and *de novo* assembly. Because the eukaryotic genomes (especially for mammals) usually contain many repetitive and homologous regions, these elements may result in the alignment ambiguities and assembly collapse. The sequencing errors of NGS further aggravate the difficulties for precisely conducing short-read alignment, *de novo* assembly, SNP calling, RNA-editing identification, gene fusion detection and other NGS applications. To alleviate the severity of these NGS problems, corresponding analytical algorithms are undergoing fast revolution and an increasing number of related software are emerging to address these challenges. In this issue, Hong and his colleagues have critically reviewed the related bioinformatic tools in each step of NGS data generation and processing [25]. Powerful bioinformatics tools can solve a portion of NGS problems to some extent, but the improvements of sequencing technologies are the fundamental way to thoroughly conquer the NGS limitations.

To conduct the NGS data analyses, lots of bioinformatics algorithms have been developed for diverse applications. Short-read alignment and assembly are two basic analyzing approaches, and they are crucial for reference-based and reference-independent studies, respectively [11,12]. In general, when the reference genome is available, the first step is to implement short-read mapping; otherwise, *de novo* assembly of short reads may need to be carried out instead. Before conducting any type of NGS data analysis, it is important to check the sequencing quality of data and remove those low quality reads or bases if necessary. Only with the high quality of the sequencing data, can we correctly analyze them and obtain accurate results. For each kind of NGS applications in genomics, transcriptomics and epigenomics, a number of relevant tools may be available to accomplish the associated analysis. Meanwhile, each one of the software often has multiple parameters that can be set by users based on their requirements. Each software has its own strengths and weaknesses, moreover, distinct parameter settings of the same tool may also result in different results.

Therefore, it is necessary to test the performance and parameters of corresponding tools before employing appropriate software on the basis of sequencing data properties. To this end, two papers in this issue have evaluated the performance of different bioinformatics tools for transcriptome reconstruction with RNA-Seq data [26,27]. On the other hand, multiple analyzing steps are normally needed to reach the research goals, and the tools from each step are usually needed to combine to finish all the involved analyses. Thus, choosing an optimal combination of bioinformatics software is high recommended to fulfill the aims of the study as well.

The NGS technologies are in continuous revolution along with the improvements of correlated bioinformatics algorithms. Many aspects of NGS technologies will be significantly improved in the future, including the short read length, sequencing depth, relatively higher sequencing errors, the sequencing cost and time [28]. Those progresses will vastly facilitate the applications of NGS technologies in personalized medicine field. New powerful bioinformatics tools will also be designed to meet the changes of NGS technologies and effectively handle the NGS data with higher accuracy. It is worth noting that the human reference genome and related gene annotations are crucial for the application of personalized medicine, however, it is still incomplete [29]. With the advances in both sequencing technologies and computational algorithms, the reference genome will be further refined and those unidentified genes/isoforms will be eventually annotated. Collectively, the NGS technologies and relevant bioinformatics tools provide us important approaches to explore masses of biological issues related to human diseases and drug sensitivities and illustrate their underlying mechanisms, but we need to carefully process and interpret the NGS data in consideration of the limitations from both technologies and algorithms. We believe the evolutions of sequencing technologies and analyzing tools will consistently boost the developments and realization of personalized medicine and benefit the health of human beings, the future of the NGS technologies on personalized medicine looks very promising but with great challenges.

*This work was supported by the National Basic Research Program of China (2010CB945401), the National Natural Science Foundation of China (31240038) and Graduate School of East China Normal University.*

1 Metzker M L. Sequencing technologies—the next generation. Nat Rev Genet, 2010, 11: 31–46
2 Teer J K, Mullikin J C. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet, 2010, 19: R145–151
3 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
4 Ozsolak F, Milos P M. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet, 2011, 12: 87–98
5 Harris R A, Wang T, Coarfa C, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol, 2010, 28: 1097–1105
6 Fouse S D, Nagarajan R O, Costello J F. Genome-scale DNA methylation analysis. Epigenomics, 2010, 2: 105–117
7 Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. Cell, 2007, 129: 823–837
8 Zhou X, Ren L, Li Y, et al. The next-generation sequencing technology: a technology review and future perspective. Sci China Life Sci, 2010, 53: 44–57
9 Wu J, Xiao J, Zhang R, et al. DNA sequencing leads to genomics progress in China. Sci China Life Sci, 2011, 54: 290–292
10 Jiang T, Yang L, Jiang H, et al. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. Sci China Life Sci, 2011, 54: 945–952
11 Chen G, Wang C, Shi T. Overview of available methods for diverse RNA-Seq data analyses. Sci China Life Sci, 2011, 54: 1121–1128
12 Chen G, Yin K, Wang C, et al. *De novo* transcriptome assembly of RNA-Seq reads with different strategies. Sci China Life Sci, 2011, 54: 1129–1133
13 Yuan L, Ren L, Li Y, et al. A complete genome assembly of *Glaciecola mesophila* sp. nov. sequenced by using BIGIS-4 sequencer system. Sci China Life Sci, 2011, 54: 835–840
14 Hao D, Ma P, Mu J, et al. *De novo* characterization of the root transcriptome of a traditional Chinese medicinal plant *Polygonum cuspidatum*. Sci China Life Sci, 2012, 55: 452–466
15 Gong W, Pan L, Lin Q, et al. Transcriptome profiling of the developing postnatal mouse testis using next-generation sequencing. Sci China Life Sci, 2013, 1: 1–12
16 Chen G, Yin K, Shi L, et al. Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. PLoS One, 2011, 6: e28318
17 Hou R, Yang Z, Li M, et al. Impact of the next-generation sequencing data depth on various biological result inferences. Sci China Life Sci, 2013, 56: 104–109
18 Jiang L C, Schlesinger F, Davis C A, et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res, 2011, 21: 1543–1551
19 Qin T, Yu Y, Du T, et al. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci, 2013, 56: 134–142
20 Wadelius M, Alfirevic A. Pharmacogenomics and personalized medicine: the plunge into next-generation sequencing. Genome Med, 2011, 3: 78
21 Toma I, St Laurent G, McCaffrey T A. Toward knowing the whole human: next-generation sequencing for personalized medicine. Pers Med, 2011, 8: 483–491
22 Zaneveld J, Wang F, Wang X, et al. Dawn of ocular gene therapy: implications for molecular diagnosis in retinal disease. Sci China Life Sci, 2013, 56: 125–133
23 Guo Y, Shi L, Hong H, et al. Studies on abacavir-induced hypersensitivity reaction: a successful example from translational pharmacogenetics to personalized medicine. Sci China Life Sci, 2013, 56: 119–124
24 National Institute of Health. THE $1000 GENOME. http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-04-003.html, http://www.genome.gov/10000368
25 Hong H, Zhang W, Shen J, et al. Critical role of bioinformatics in translating huge amounts of next-generation sequencing data into personalized medicine. Sci China Life Sci, 2013, 56: 110–118
26 Lu B, Chen G, Shi T. Comparisons of transcriptome reconstruction methods for RNA-seq data. Sci China Life Sci, 2013, 56: 143–155
27 Clarke K, Yang Y, Marsh R, et al. Comparative analysis of *de novo* transcriptome assembly. Sci China Life Sci, 2013, 56: 156–162
28 Kedes L, Campany G. The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE Competition. Nat Genet, 2011, 43: 1055–1058
29 Chen G, Li R, Shi L, et al. Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics, 2011, 12: 590